Real-Time 3D Motion Capture by Monocular Vision and Virtual Rendering

David Antonio Gómez Jáuregui and Patrick Horain

Institut Mines-Télécom, Télécom SudParis, 9 rue Charles Fourier,
91011 Evry Cedex, France
David_Gomez1380@yahoo.com.mx, Patrick.Horain@Telecom-Sudparis.eu

Abstract. Avatars in networked 3D virtual environments allow users to interact over the Internet and to get some feeling of virtual telepresence. However, avatar control may be tedious. Motion capture systems based on 3D sensors have recently reached the consumer market, but webcams and camera-phones are more widespread and cheaper. The proposed demonstration aims at animating a user's avatar from real time 3D motion capture by monoscopic computer vision, thus allowing virtual telepresence to anyone using a personal computer with a webcam or a camera-phone. This kind of immersion allows new gesture-based communication channels to be opened in a virtual inhabited 3D space.

Keywords: 3D motion capture, monocular vision, 3D/2D registration, particle filtering, real-time computer vision.

1 Introduction

Avatars in 3D virtual environments allow to enhance remote users' mutual perception by mimicking their motion in real-time. This contributes to gesture-based communication in a virtual inhabited 3D environment [1].

Such an interface relies on real-time user-friendly motion capture. Classical sensors (e.g. data gloves, magnetic sensors or optical markers) require to be attached on the performer's body and limbs. Instead, computer vision allows an inexpensive and practical approach [2], [3]. We focus on 3D human motion capture in real-time without markers. It is still a challenge because of the ambiguities resulting of the lack of depth information, of possible partial occlusion of the body limbs, of the high number of degrees of freedom and of the varying cloth of a person. We focus on 3D human motion capture in real-time without markers. It is still a challenge because of the ambiguities resulting of the lack of depth information, of possible partial occlusion of the body limbs, of the high number of degrees of freedom and of the varying cloth of a person. Webcams are very low-cost and widespread consumer devices, in contrast to recent 3D sensors (e.g. time-of-flight cameras [4] or Kinect [5]) which are non-standard extensions to personal computers. Even mobile devices can capture a video stream and have it sent (through the Internet) and processed on a remote computer.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 663–666, 2012.

In this demonstration, we animate a user's avatar from real-time 3D motion capture by monoscopic vision with consumer hardware. We use cameras that come with many personal computers (webcams) or mobile devices (smart phones) (Figure 1).

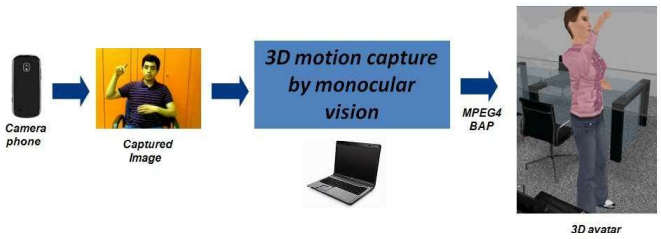

Fig. 1. Animating a 3D avatar from real-time 3D motion capture by monocular vision with consumer hardware

2 Real-Time Model-Based 3D Motion Capture

Our system works by registering a 3D articulated upper-body model to an input video stream. A vector of position and joint angles parameters fully defines the model pose. Registration consists in searching for the best match, with respect to those parameters, between primitives extracted from the 3D model and those from the captured image (Figure 2).

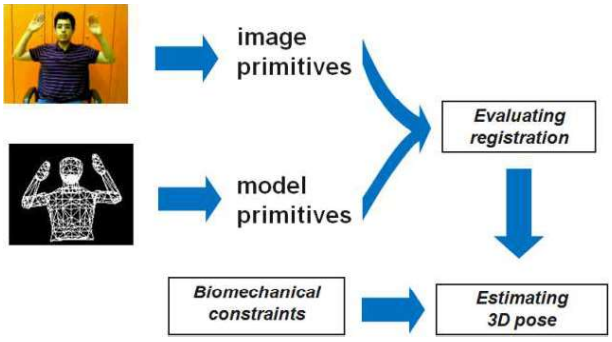

Fig. 2. The implemented approach for 3D motion capture

Our motion capture system is divided into two main stages: initialization and tracking. In the first step, the appearance of the background and the user are learnt, and registration is used to adjust the morphology of the 3D model to the user.

In the tracking step, the pose of the 3D body model that optimally matches the input image is searched iteratively. Biomechanical constraints allow to discard poses that are physically unreachable by the human body [6]. At each video

frame, the iterative registration process is initialized at the pose estimated at the previous frames. We extract image features (color regions and edges) from the input video sequence and estimate the 3D pose that best matches them in real-time [6]. Monocular ambiguities due to the lack of depth information are handled in a particle filter framework enhanced with heuristics, under the constraint of real-time computation [7]. Here, we have developed a more sophisticated particle filter algorithm to reduce the number of particles (3D pose hypotheses) required for monocular 3D motion capture. It integrates a number of heuristics and search strategies into the CONDENSATION [8] approach to guide particles toward highly probable solutions. First, in the resampling step, children particles are selected and grouped according to their parents weights using a weight-based resampling heuristic. Then, in the prediction step, large groups of particles are guided toward maximums of the posterior density using local optimization while small group of particles are diffused randomly in the pose space. Ambiguities from monocular images are handled by computing new samples by kinematic flipping [9]. A hierarchical partitioned sampling is used to diffuse particles more efficiently based on motion observations. 3D poses are described using end-effector position to better model uncertainty in depth direction. Finally, evaluation of particles is accelerated by a parallelized GPU implementation. Our real-time particle filter algorithm that combines all the previous heuristics did significantly improved the tracking robustness and accuracy using as little as 200 particles in 20 degrees of freedom state space. Finally, the best particle (3D pose) obtained at each image is encoded in the MPEG-4 BAP format [10]. BAP parameters are sent through a TCP/IP socket connection to a local or remote computer where the 3D collaborative virtual environment application is installed (Figure 3).

Fig. 3. Real-time 3D motion capture by computer vision and virtual rendering. For each captured image (top left inlays), the 3D model is projected with the pose that best matches the primitives (bottom left inlays). The virtual avatar (OpenSpace3D [11]) renders the captured gestures.

3 Conclusion and Future Work

We have developed an inexpensive system for robust real-time upper-body motion capture from monocular images using consumer hardware with a single camera. This system can be used for gestural communication in a 3D virtual environment in order to reinforce user interaction and the feeling of telepresence. Other potential applications include gesture-based human-computer interaction for networked virtual environments, home video monitoring of fragile elder people, human-robot interaction, multi-modal interfaces, etc.

References

1. Horain, P., Soares, J.M., Kumar, P., Bideau, A.: Virtually enhancing the perception of user actions. In: 15th International Conference on Artificial Reality and Telexistence (ICAT 2005), Christchurch, New Zealand, pp. 245–246 (2005)
2. Moeslund, T., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
3. Poppe, R.W.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108, 4–18 (2007)
4. Lindner, M., Schiller, I., Kolb, A., Koch, R.: Time-of-flight sensor calibration for accurate range sensing. Computer Vision and Image Understanding 114, 1318–1328 (2010)
5. Microsoft: Kinect - xbox.com (2011)
6. Jáuregui, D.A.G., Horain, P.: Region-Based *vs.* Edge-Based Registration for 3D Motion Capture by Real Time Monoscopic Vision. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2009. LNCS, vol. 5496, pp. 344–355. Springer, Heidelberg (2009)
7. Jáuregui, D.A.G., Horain, P., Rajagopal, M.K., Karri, S.S.K.: Real-time particle filtering with heuristics for 3D motion capture by monocular vision. In: Proceedings of the 2010 IEEE International Workshop on Multimedia Signal Processing (MMSP 2010), Saint-Malo, France (2010)
8. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. IJCV: International Journal of Computer Vision 29, 5–28 (1998)
9. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3D human tracking. In: International Conference on Computer Vision and Pattern Recognition, Madison, WI, pp. 69–76 (2003)
10. ISO/IEC 14996-2: Information technology-coding of audio-visual objects-part 2: visual (2001)
11. I-Maginer: Open source platform for 3D environments (2010)